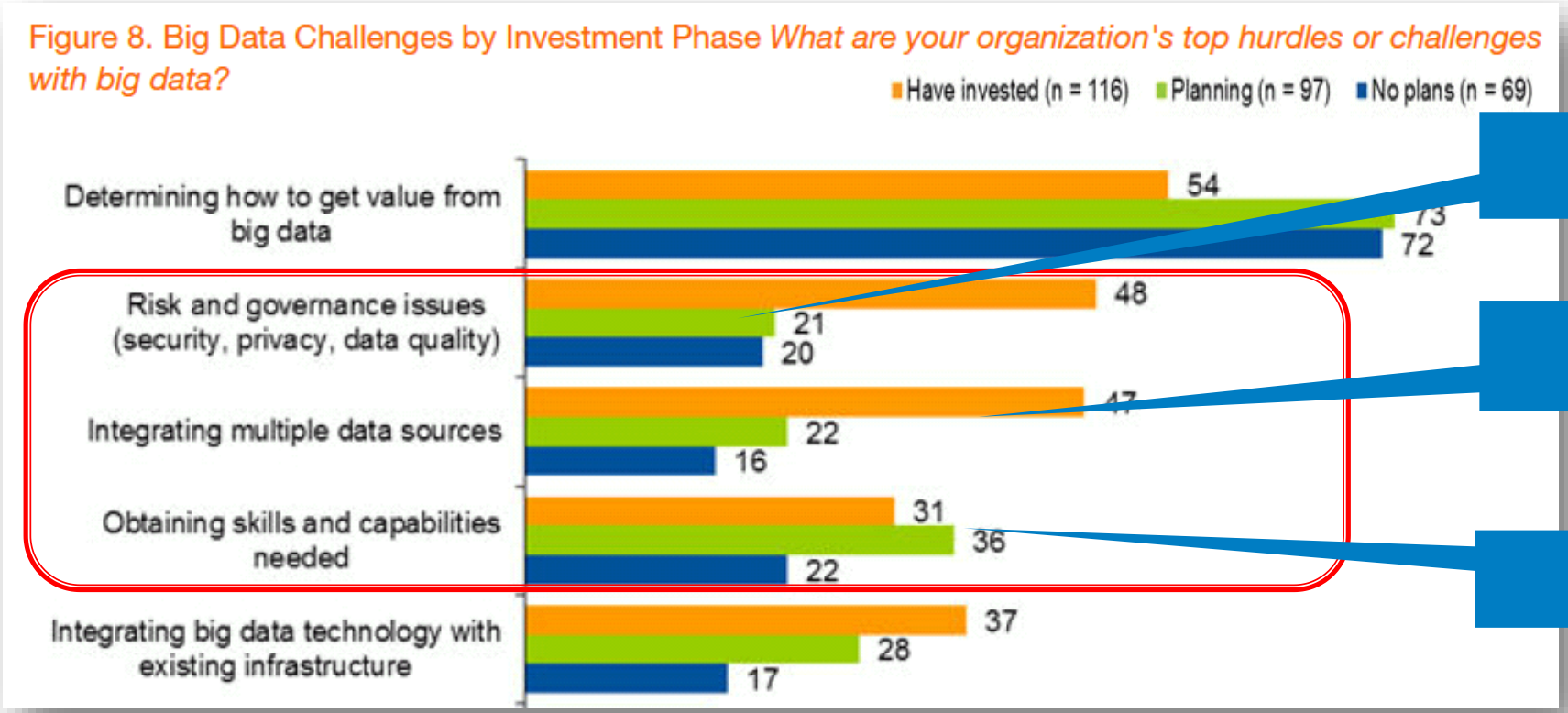
SAS® FORUM POLSKA 2015

SAS Access to Hadoop, SAS Data Loader for Hadoop
Integracja środowisk SAS i Hadoop

Piotr Borowik

sas THE POWER TO KNOW.

Wyzwania związane z Big Data



Jakość danych

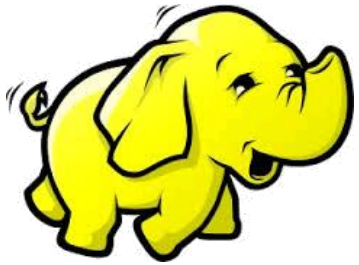
Integracja danych

Brak komptenecji

Source: **Gartner** (Sep 2014), **Big Data Investment Grows but Deployments Remain Scarce in 2014** *By Nick Heudecker, Lisa Kart*

SAS & Intel Study

SAS razem z Intel zapytali 300 menadżerów IT największych firm skandynawskich o ich podejście do Hadoop
http://nordichadoopsurvey.com



Główny powód dla użycia Hadoop

60% - zaawansowana analityka, data discovery, analytical lab

22% - przyspieszenie przetwarzań

Adopcja / Problemy

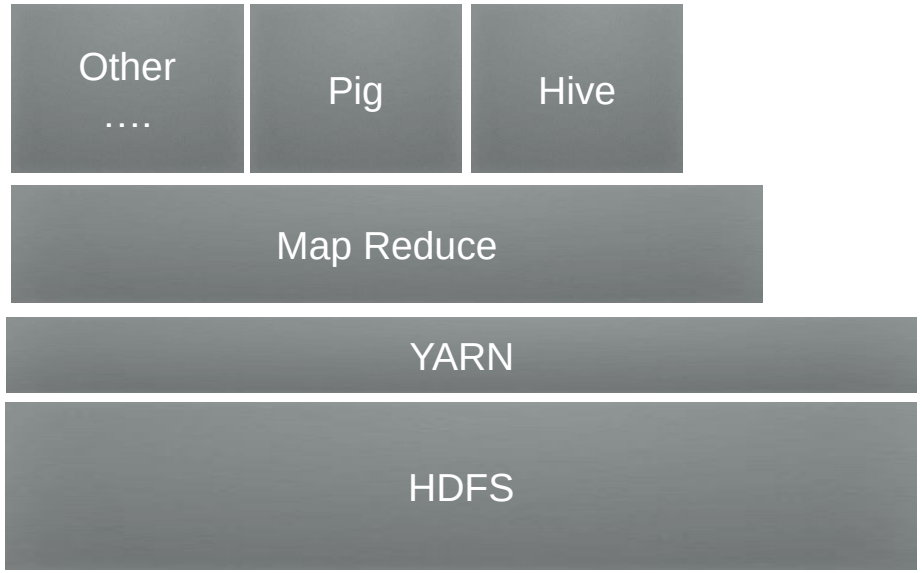
35% - brak zasobów i kompetencji

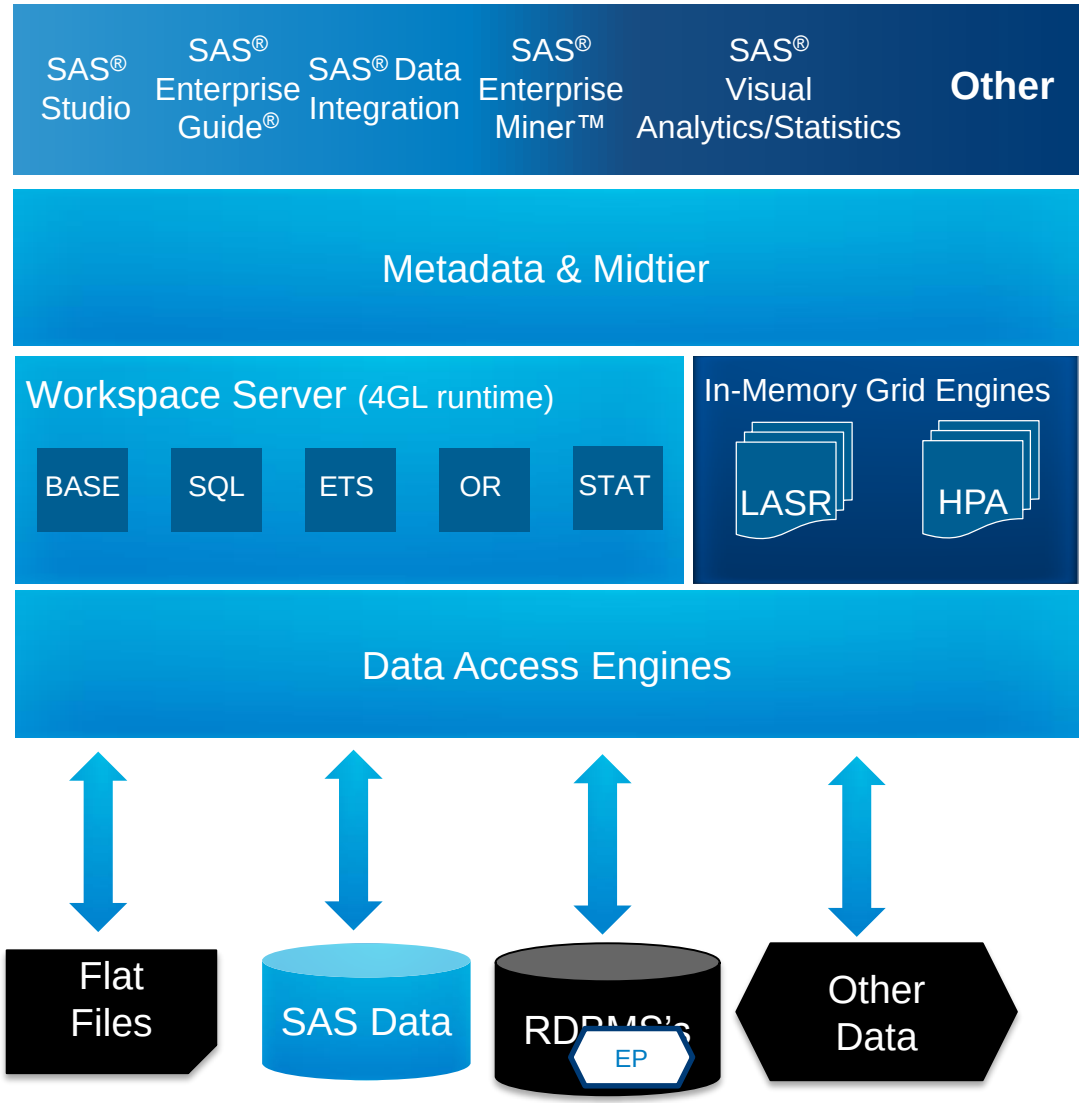
SAS vs Hadoop

End User / Business Applications

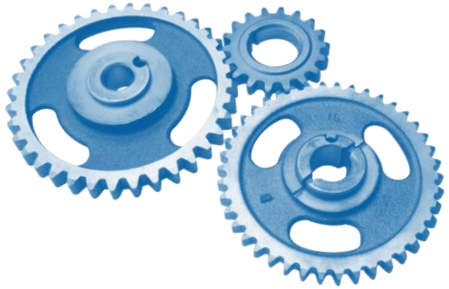
Data Processing

Permanent File Store





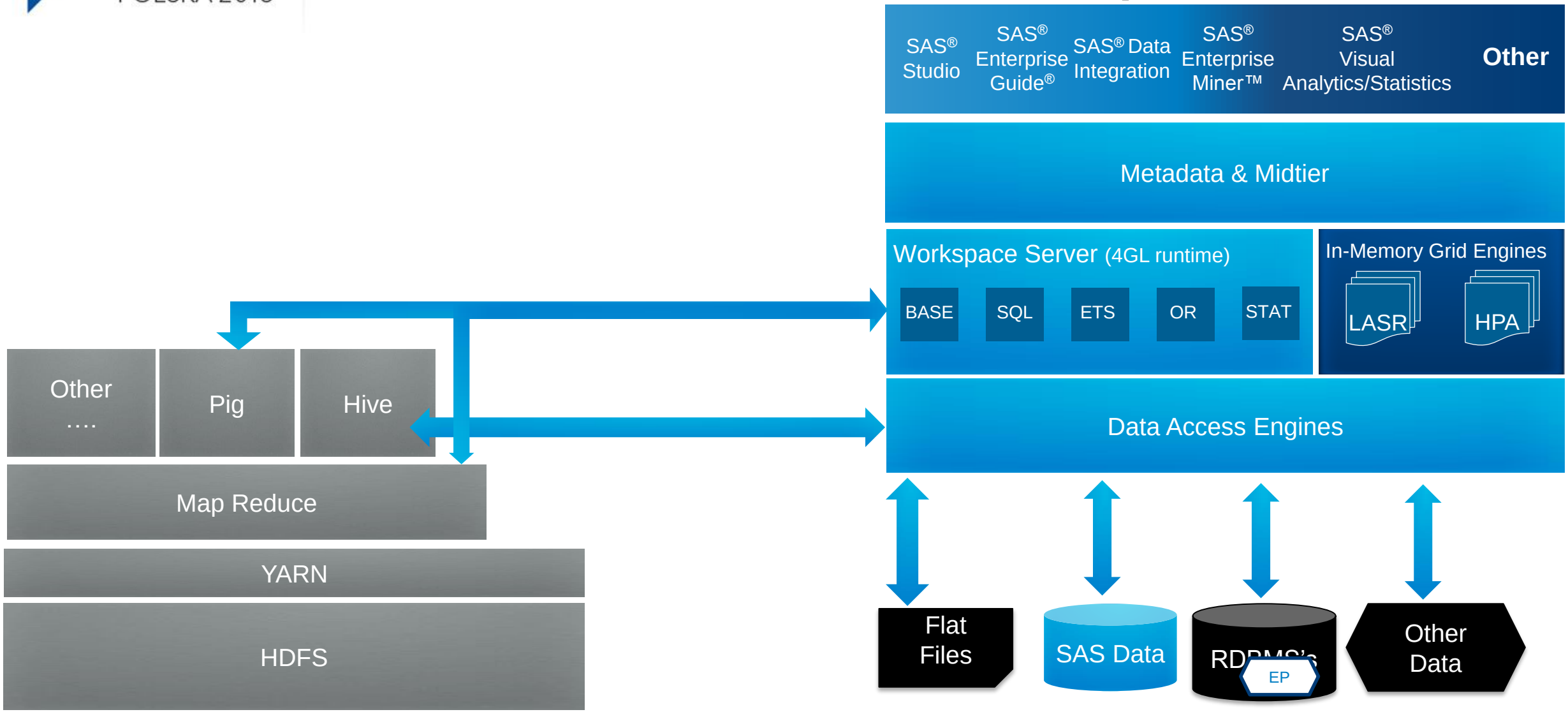
SAS Access to Hadoop, HAWQ, Impala,…



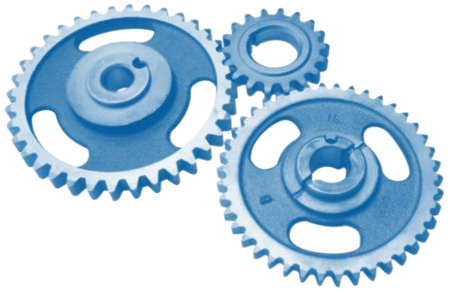
Natywne mechanizmy integracji z repozytorium danych w środowisku SAS

- Mechanizm Libname – czyli dane widoczne jak tabele SAS
- Implicit pass-through – automatyczne generowanie SQL bazy danych, w tym konwersja PROC SQL
- Explicit pass-through – jawny SQL w dialekcie bazy
- Automatyczna generacja SQL dla niektórych procedur
- Optymalizacja (wybrane przykłady):
  - Wielowątkowy odczyt i autopartycjonowanie
  - Natywne mechanizmy ładowania
  - Przekazywanie zapytań (np. JOIN) do bazy
  - Heterogeniczny JOIN z optymalizacją
  - Przetwarzanie WHERE w procedurach SAS

Scenariusz 1. SAS klientem hadoop



Embedded Process - SAS in-database



EP udostępnia run-time SAS (tzw. TK) w „platformach MPP”

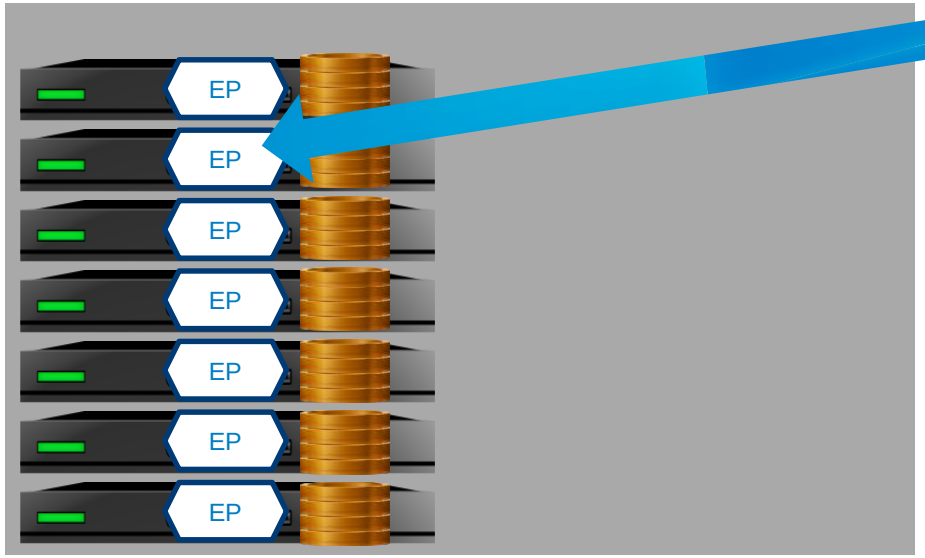
- Wykonuje program w języku SAS DS2
- Implementacja specyficzna (zoptymalizowana) dla danego środowiska MPP (skalowalność taka jak środowiska)
- Wykorzystuje rozproszoną architekturę Hadoop (Map-Reduce) lub mechanizmy platformy MPP (integracja poprzez Table Function UDF) dla optymalizacji i przetwarzania równoległego

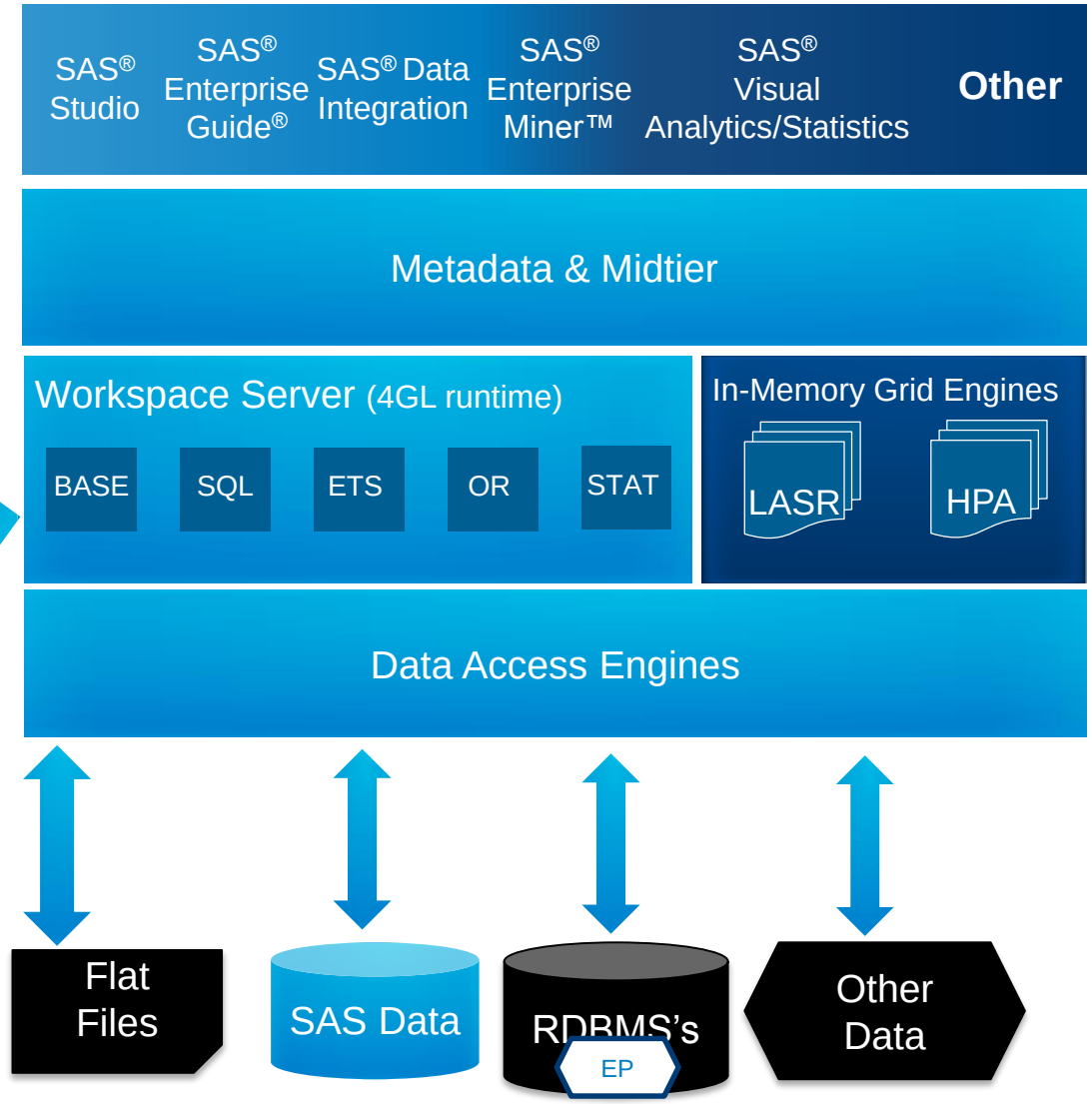
Jest zarządzany przez mechanizm workload management platformy MPP
(np. przez YARN – jak każda aplikacja M-R)

W środowisku Hadoop wykorzystuje Hive, HCatalog oraz natywny dostęp do HDFS aby zapisywać/odczytywać dane (Avro, ORC, Parquet, RCFile, sequence, binary, delimited, XML)

Scenariusz 2. SAS in-database

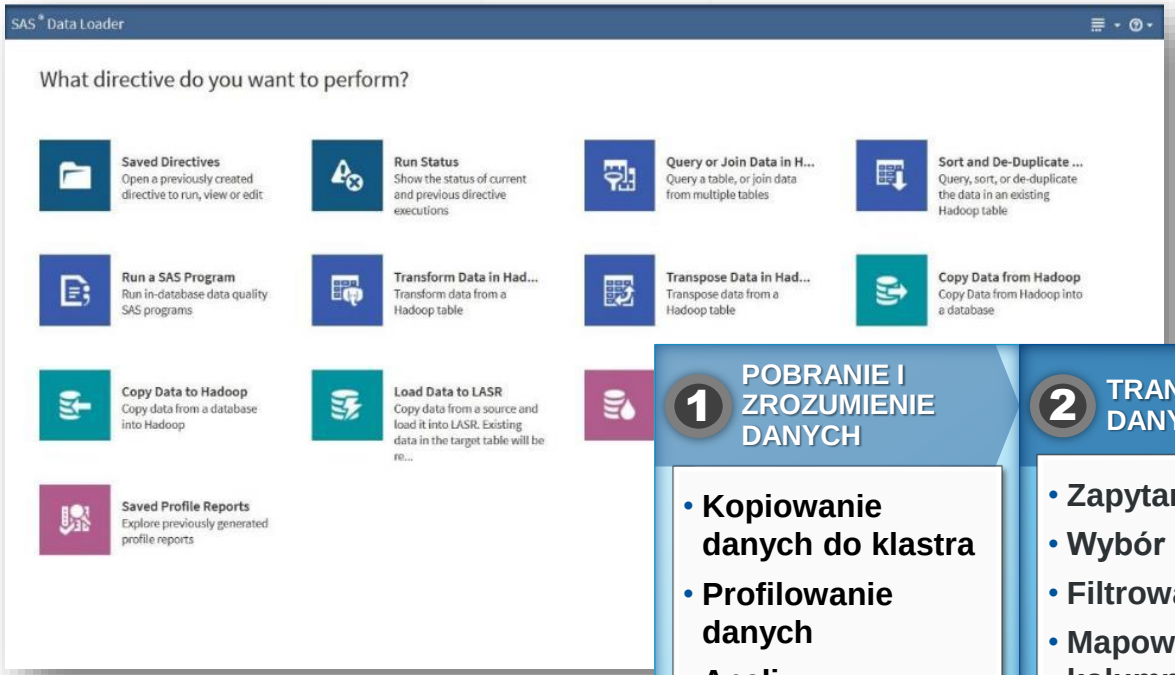
- In-Database Scoring
- DS2 Processing
- Data Step accelerator
- DQ Processing





SAS® FORUM POLSKA 2015

SAS DATA LOADER FOR HADOOP



1 POBRANIE I ZROZUMIENIE DANYCH

- Kopiowanie danych do klastra
- Profilowanie danych
- Analizy identyfikacyjne
- Zapytania
- Import plików lokalnych

2 TRANSFORMACJA DANYCH

- Zapytania
- Wybór kolumn
- Filtrowanie
- Mapowanie i kolumny wyliczane
- Sortowanie / Aggregacja
- Transpozycja
- Transformacje danych
- Usuwanie wierszy

3 CZYSZCZENIE DANYCH

- Walidacja
- Parsowanie
- Standardyzacja
- Zamiana wielkości liter
- Analiza płci
- Analiza wzorców
- Ekstrakcja pól

4 INTEGRACJA DANYCH

- Łączenie danych
- Tworzenie Match kodów
- Sortowanie i deduplikacja
- Agregacja
- Uruchamianie kodów SAS 4GL
- Uruchamianie kodów Hive

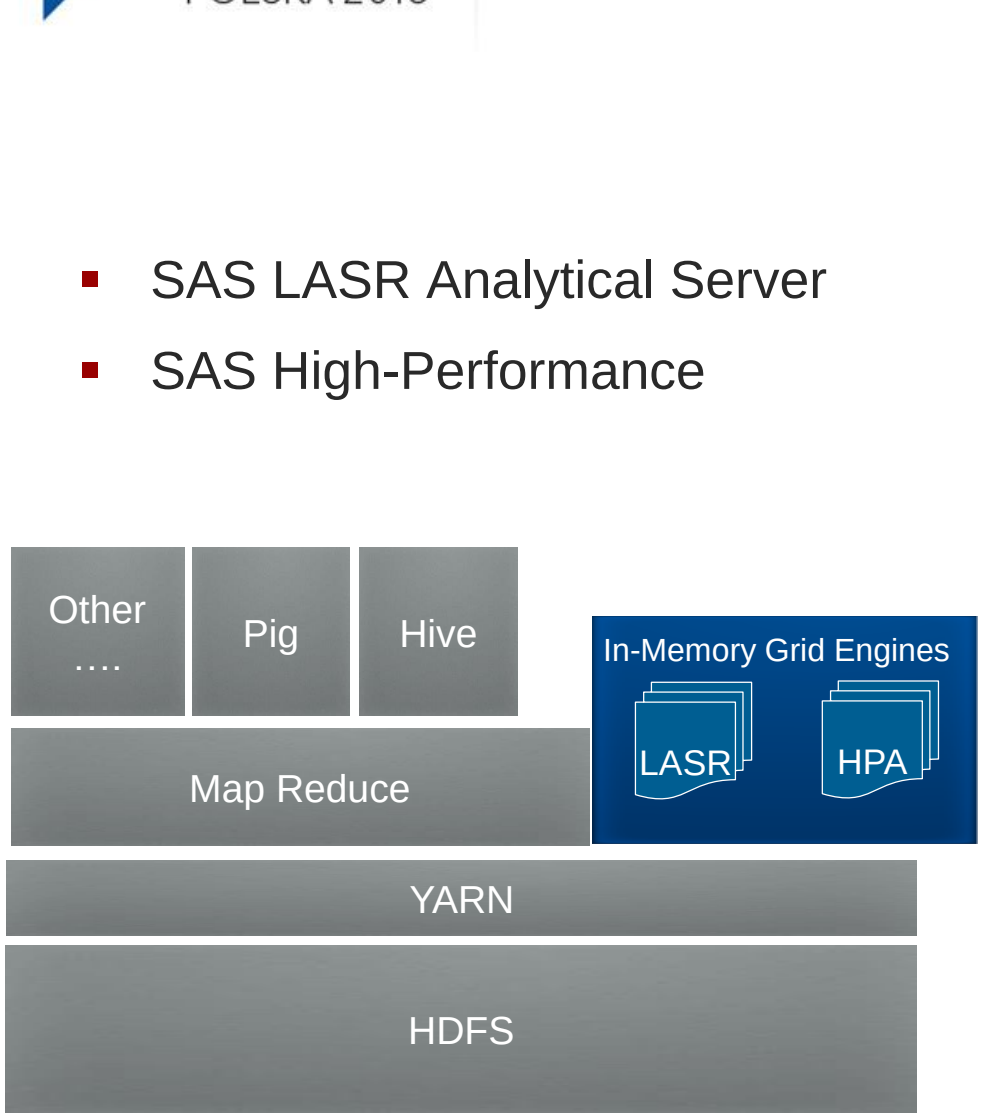
5 DOSTARCZENIE DANYCH

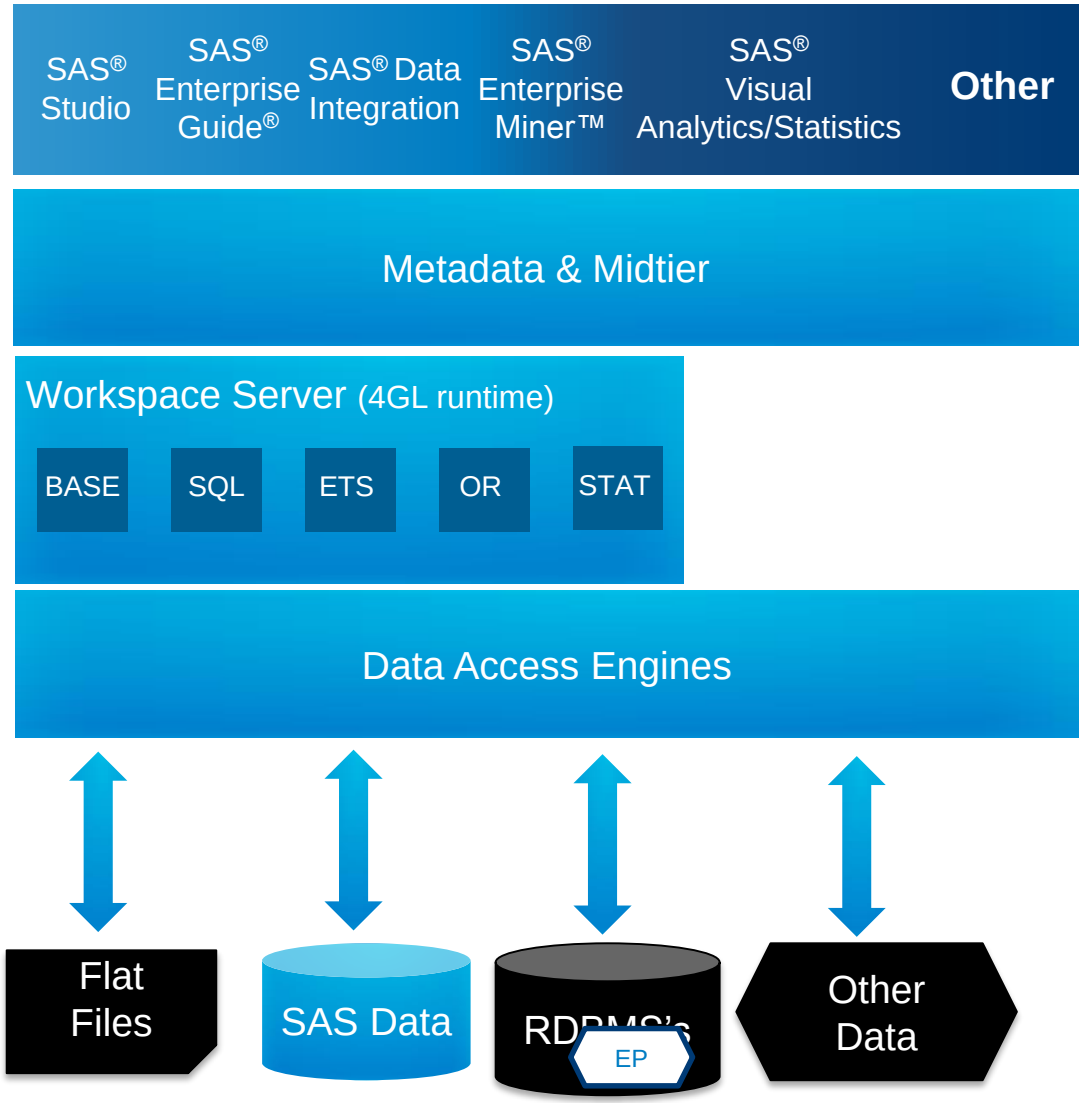
- Ładowanie do SAS LASR Server
- Tworzenie tabel
- Tworzenie perspektyw/view
- Kopia do RDBMS

SAS® FORUM POLSKA 2015

Scenariusz 3. SAS razem z hadoop

- SAS LASR Analytical Server
- SAS High-Performance





sas THE POWER TO KNOW.

SAS® FORUM
POLSKA 2015

sas THE POWER TO KNOW.